Un nuevo modelo de aprendizaje para el estudio de secuencias de símbolos

José Luis Triviño-Rodriguez, Rafael Morales-Bueno

Departamento de Lenguajes y Ciencias de la Computación (Universidad de Málaga)
E.T.S.I. Informática (Campus Teatinos)
Málaga, 29071
{trivino,morales}@lcc.uma.es

Resumen

Esta tesis presenta un novedoso modelo para el estudio de la dinámica temporal de un sistema basado en las cadenas de Markov de orden variable. Este modelo recibe el nombre de *Grafo Sufijo de Predicción Multiatributo* (MPSG). El enfoque principal de este modelo se basa en representar el estado del sistema por medio de la combinación de los valores de diferentes atributos o características del mismo en lugar de utilizar un único símbolo al estilo de las cadenas de Markov.

Palabras clave: Cadenas de Markov, MPSG, PSA, PST, Secuencias de símbolos.

1. Introducción

De entre los diferentes modelos desarrollados en Aprendizaje Automático para el estudio de la evolución temporal de un sistema, destacan especialmente aquellos que representan el estado del sistema en cada instante de tiempo mediante un símbolo. Por tanto, este tipo de modelos representa la evolución temporal del sistema mediante una secuencia de símbolos.

Uno de los modelos más extendidos que utiliza este enfoque para la representación del problema son las *cadenas de Markov* [8], [12] y modelos derivados de las mismas [2], [14], [4], [5]. Estos modelos, llamados estocásticos, analizan el comportamiento temporal de un sistema a través del cálculo de la distribución de probabilidades del siguiente símbolo en la secuencia condicionada por los símbolos que le preceden.

Esta tesis presenta un novedoso modelo para el estudio de la dinámica temporal de un sistema basado en las cadenas de Markov de orden variable. Este modelo recibe el nombre de *Grafo Sufijo de Predicción Multiatributo* (MPSG). El enfoque principal de este modelo se basa en representar el estado del sistema por medio de la combinación de los valores de diferentes atributos o características del mismo en lugar de utilizar un único símbolo al estilo de las cadenas de Markov. De esta forma, la evolución temporal del sistema se representa en este modelo en forma de un conjunto de secuencias de símbolos que evolucionan de forma paralela. Cada secuencia representa la evolución temporal de un atributo del sistema. Además, la longitud de memoria necesaria para calcular la distribución de probabilidades del siguiente símbolo es variable e independiente para cada atributo. De esta forma, este modelo puede entenderse como una extensión de las cadenas de Markov de longitud variable a un sistema multiatributo.

Debido a la representación utilizada, el modelo

MPSG combina el análisis de la evolución temporal del sistema con el estudio de las relaciones causa/efecto entre los diferentes atributos del mismo. Por tanto, este modelo puede ser considerado también como una extensión de la dimensión temporal de los árboles de decisión.

Basándonos en lo anterior, se puede afirmar que el modelo MPSG constituye una unificación de dos de los más importantes modelos existentes: las cadenas de Markov y los árboles de decisión [11].

Respecto a la aplicabilidad práctica de modelo MPSG, considerando la expresividad del modelo de hipótesis elegido, es fácil constatar la existencia de numerosos problemas a los que puede ser aplicado. De esta forma, este modelo resulta especialmente adecuado para el estudio de la evolución temporal de sistemas cuyo estado venga expresado por la combinación de diferentes parámetros del mismo. De forma especial, el modelo MPSG será adecuado para añadir la dimensión temporal al análisis realizado de un sistema por medio de árboles de decisión.

Con objeto de mostrar experimentalmente el funcionamiento del modelo se describen dos aplicaciones prácticas del modelo MPSG: el modelado y generación de música y el análisis morfológico.

En la primera de las aplicaciones, el modelo MPSG es utilizado para crear un modelo de las corales de J.S. Bach [7]. Este modelo ha sido utilizado posteriormente para generar piezas inéditas. Utilizando diversos test estadísticos se ha demostrado que las piezas generadas siguen la misma distribución de probabilidades que las corales iniciales. Además, estas piezas han sido utilizadas en un test de audición donde se ha podido constatar que los sujetos a los que se les practicó el test no eran capaces de distinguir las piezas originales de las generadas.

Respecto a la aplicación del modelo MPSG al análisis morfológico, se ha desarrollado un etiquetador de español basado en este modelo. Este etiquetador ha sido entrenado y evaluado utilizando el corpus LexEsp de la UPC, obteniendose resultados similares al de los mejores etiquetadores existentes en la actualidad.

2. Contribución

La contribución de esta tésis se centra en dos aspectos fundamentales: por un lado, el desarrollo de nuevos modelos teóricos de aprendizaje de secuencias de símbolos y su aplicación práctica; por otro lado, la comparación a nivel teórico de modelos de aprendizaje y la generalización desarrollada a través del modelo MPSG de los árboles sufijos de predicció (PST) y los árboles de desición (TDIDT).

En la línea de desarrollo de nuevos modelos de modelado y aprendizaje de secuencias de símbolos, el núcleo de esta tésis se centra en el desarrollo de un nuevo modelo llamado *Grafo Sufijo de Predicción Multiatributo* (MPSG).

El modelo MPSG ha sido desarrollado como una generalización multiatributo del modelo desarrollado por Dana Ron llamado *Sufijo de Predicción* (PST)[13]. El objetivo de este desarrollo teórico ha sido el de poder modelar de forma eficiente la evolución temporal de sistemas expresados en forma de secuencias paralelas de símbolos. La aplicación práctica de este modelo se ha puesto de manifiesto en áreas tan diferentes como el modelado y generación de música [1][15] como en el etiquetado de partes de la oración [3], [6], [9], [16].

Otra de las aportaciones de esta tésis tiene su origen en el estudio teórico llevado a cabo sobre el modelo MPSG desarrollado. Este estudio a puesto de maniefiesto como el modelo MPSG puede ser considerado como una generacilazión de dos modelos ampliamente extendidos: las cadenas de Markov y los árboles de decisión [10]. En este sentido, el modelo MPSG puede considerarse como la extensión multiatributo de las cadenas de Markov o, por otro lado, como la inclusión de la evolución temporal del sistema en el modelado realizado por un árbol de desición [17].

3. Estructura de la memoria

El núcleo de la tesis es el modelo de *Grafo Sufijo de Predicción Multiatributo (MPSG)*: definición, propiedades, algoritmo de aprendizaje y aplicaciones. Previamente, el primer capítulo de esta tésis está dedicado al modelado del tiempo en la Inteligencia Artificial. Este capítulo describe los diferentes modelos utilizados para la representación del tiempo a lo largo de la historia de la Inteligencia Artificial y las características de los

mismos.

Posteriormente, el segundo capítulo describe cómo los diferentes modelos de aprendizaje automático existentes modelan el tiempo y la evolución en el mismo de un sistema. Para ello se analizaran los diferentes modelos existentes desde el modelo de descubrimiento de conocimiento basado en eventos hasta los modelos de aprendizaje automático usados habitualmente para un estudio estático de problemas pero en los cuales se puede incluir la dimensión temporal. Especial atención se presta en este capítulo a los modelos derivados de las cadenas de Markov como los modelos ocultos de Markov y los automatas sufijos probabilísticos (PSAs).

El siguiente capítulo (capítulo tres) define formalmente el modelo de grafo sufijo de predicción multiatributo desarrollado. Además, en este capítulo se analiza y demostra la equivalencia entre los MPSGs y los arboles de decisión.

En el cuarto capítulo se describe el modelo de aprendizaje de MPSGs y, más concretamente, se analizan los algoritmos de aprendizaje de MPSGs. Estos algoritmos permiten calcular, a partir de una muestra suficientemente grande un MPSG que define una distribución de probabilidades similar al MPSG que generó la muestra. Además se demostra que es suficiente con una muestra de tamaño polinómico respecto a los parámetros del MPSG para realizar el entrenamiento.

Una vez descritos los modelos teóricos y algoritmos de aprendizaje, el quinto capítulo describe la aplicación práctica de este modelo. Concretamente este capítulo muestra dos aplicaciones reales del modelo: la desambiguación morfológica y la predicción y generación de música.

Por último, el sexto capítulo desarrolla una serie de conclusiones alcanzadas a lo largo del desarrollo del modelo de MPSG. Además, este capítulo describe diferentes líneas de desarrollo e investigación basadas en los MPSGs.

4. Conclusiones

En esta tésis se ha desarrollado y estudiado tanto teoricamente como en su aplicación práctica un nuevo modelo de aprendizaje llamada Grafo Sufijo de Predicción Multiatributo (MPSG). Las características de este modelo se pueden resumir en tres puntos fundamentales:

1. Los MPSGs modelan sistemas cuya evolución en el tiempo se expresa en forma de secuencias de símbolos paralelas entre sí.

2. La longitud de memoria para cada atributo (secuencia) es variable e independiente de la longitud de memoria del resto de los atributos.

3. El modelo MPSG calcula las secuencias mínimas de símbolos para cada atributo necesarias para determinar la distribución de probabilidad del siguiente símbolo del atributo objetivo. Ello permite, en ciertos casos, generar una descripción del problema más concisa que un árbol de decisión.

Por otro lado, la aplicación práctica de este modelo se ha puesto de manifiesto en dos aplicaciones descritas en la tésis: etiquetado de partes de la oración y modelado y generación de música.

En conclusión, se puede considerar a los MPSGs un nuevo modelo de aprendizaje de secuencias de símbolos que combina varias de las características de modelos existentes mantiendo un nivel de eficiencia en el entrenamiento aceptable. Si bien no está pensado para sustituir a otros modelos de aprendizaje, sus características lo hacen adecuado para ser aplicado con éxito en numerosos problemas reales mejorando incluso los resultados obtenidos por estos modelos.

Referencias

[1] G. Assayag, S. Dubnov, and O. Delerue. Guessing the composer's mind: Applying universal prediction to musical style. In *Proceedings of the ICMC*, 1999.

[2] P. Buhlmann. Model selection for variable length markov chains and tuning the context algorithm. Technical report, ETH Zurich, 1997.

[3] Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. Equations for part-ofspeech tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 784–789, 1993.

[4] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical Hidden Markov Model: Analysis and applications. *Machine Learning*, 32:41, 1998.

[5] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–273, 1997.

[6] Fred Jelinek. Robust part-of-speech tagging using a hidden markov model. Technical report, IBM, 1985.

[7] F.D. Mainous and R.W. Ottman. *Three hundred seventy-one chorales of Johann Sebastian Bach*. Harcourt Brace College Publishers, Orlando, 1966.

[8] Andrei A. Markov. Ischislenie veroiatnostei, 1924.

[9] Lluis Marquez and Horacio Rodriguez. Part-of-speech tagging using decision trees. In *European Conference on Machine Learning*, pages 25–36, 1998.

[10] J.R. Quinlan. Discovering rules by induction from large collections of examples. In D. Michie, editor, *Expert Systems in the Micho-electronic Age*. Edinburgh University Press, 1979.

[11] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[12] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE Acoustics, Speech & Signal Processing Magazine*, 3:4–16, 1986.

[13] D. Ron. *Automata Learning and its Applications*. PhD thesis, MIT, 1996.

[14] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, page 34, 1996.

[15] J.L. Triviño and R. Morales. Using multiattribute prediction suffix graphs to predict or generate music. *Computational Music Journal*, 2001. (en prensa).

[16] J.L. Triviño and R. Morales. Ussing multiattribute prediction suffix graphs for part-of-speech tagging. In *International Symposium on Intelligent Data Analysis*, Cascais (Portugal), 2001. Lecture Notes in Computer Science.

[17] J.L. Triviño-Rodriguez and R. Morales-Bueno. Grafos sufijos de predicción multiatributo. una visión unificada de las cadenas de markov y los árboles de desición. In *IX Conferencia de la Asociación Española para la Inteligencia Artificial*, volume I, pages 333–143, noviembre 2001.